Characterization of Large Structural Genetic Mosaicism in Human Autosomes

Mitchell J. Machiela,1 Weiying Zhou,1,2 Joshua N. Sampson,1 Michael C. Dean,3 Kevin B. Jacobs,4 Amanda Black,1 Louise A. Brinton,1 I-Shou Chang,5 Chu Chen,6 Constance Chen,7 Kexin Chen,8 Linda S. Cook,9 Marta Crous Bou,7,10 Immaculata De Vivo,7,10 Jennifer Doherty,11 Christine M. Friedenreich,12 Mia M. Gaudet,13 Christopher A. Haiman,14 Susan E. Hankinson,10,15 Patricia Hartge,1 Brian E. Henderson,14 Yun-Chul Hong,16 H. Dean Hosgood, III,1,7,17 Chao A. Hsiung,18 Wei Hu,1 David J. Hunter,7,10,19 Lea Jessop,21 Hee Nam Kim,20 Yeul Hong Kim,21 Young Tae Kim,22 Robert Klein,23 Peter Kraft,7 Qing Lan,1 Dongxin Lin,24,25 Jianjun Liu,26,27 Loic Le Marchand,28 Xiaolin Liang,29 Jolanta Lissowska,30 Lingeng Lu,31 Anthony M. Magliocco,32 Keitaro Matsuo,33 Sara H. Olson,29 Irene Orlow,29 Jae Yong Park,34 Loreall Pooler,35 Jennifer Prescott,7,10 Radhia Rastogi,29 Harvey A. Risch,31 Fredrick Schumacher,14 Adeline Seewo,36 Veronica Wendy Setiawan,14 Hongbing Shen,37,38 Xin Sheng,35 Min-Ho Shin,39 Xiao-Ou Shu,40 David VanDen Berg,4 Jiu-Cun Wang,11,42 Nicolas Wentzensen,1 Maria Pik Wong,43 Chen Wu,42,25 Tangchun Wu,44 Yi-Long Wu,45 Xia Huang,46,47,48 Wei Zheng,47,48 Baosen Zhou,48 Christian C. Abnet,1 Demetrus Albanes,1 Melinda C. Aldrich,47,49 Christopher Amos,50 Laufey T. Amundadottir,1 Sonja I. Berndt,4 William J. Blot,47,51 Cathryn H. Bock,52 Paige M. Bracci,53 Laurie Burdett,4,1 Julie E. Buring,54 Mary A. Butler,55 Tania Carreón,55 Nilanjan Chatterjee,1

1Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD 20892, USA; 2Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Leidos Biomedical Research Inc., Bethesda, MD 20892, USA; 3Laboratory of Experimental Immunology, Center for Cancer Research, National Cancer Institute at Frederick, NIH, Frederick, MD 21702, USA; 4BioInformed LLC, Gaithersburg, MD 20877, USA; 5National Institute of Cancer Research, National Health Research Institutes, Zhunan 35053, Taiwan, ROC; 6Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; 7Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA 02115, USA; 8Department of Epidemiology and Biostatistics, Tainan Medical University Cancer Institute and Hospital, Tainan 700040, People’s Republic of China; 9University of New Mexico, Albuquerque, NM 87131, USA; 10Channing Division of Network Medicine, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA; 11Geisel School of Medicine, Dartmouth College, Lebanon, NH 03755, USA; 12Department of Population Health Research, CancerControl Alberta, Alberta Health Services, Calgary, AB T2N 2Z9, Canada; 13Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; 14Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; 15Division of Biostatistics and Epidemiology, School of Public Health and Health Promotion, University of Massachusetts Amherst, Amherst, MA 01003, USA; 16Department of Preventive Medicine, College of Medicine, Seoul National University, Seoul 151-742, Republic of Korea; 17Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA; 18Institute of Population Health Sciences, National Health Research Institutes, Zhunan 35053, Taiwan, ROC; 19Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; 20Center for Creative Biomedical Scientists, Chonnam National University, Gwangju 500-757, Republic of Korea; 21Division of Oncology/Hematology, Department of Internal Medicine, College of Medicine, Korea University Anam Hospital, Seoul 151-742, Republic of Korea; 22Department of Oncology and Hematology, College of Medicine, University of Southern California, Los Angeles, CA 90007, USA; 23School of Public Health, University of Hawaii, Honolulu, HI 96813, USA; 24Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; 25Department of Population Health Research, CancerControl Alberta, Alberta Health Services, Calgary, AB T2N 2Z9, Canada; 26Department of Epidemiology and Biostatistics, University of Southern California, Los Angeles, CA 90033, USA; 27School of Life Sciences, Anhui Medical University, Hefei 230032, People’s Republic of China; 28Department of Epidemiology, School of Public Health, University of Southern California, Los Angeles, CA 90033, USA; 29Department of Preventive Medicine, College of Medicine, Seoul National University, Seoul 151-742, Republic of Korea; 30Department of Epidemiology and Prevention, Maria Sklodowska-Curie Cancer Center and Institute of Oncology, Warsaw 02-781, Poland; 31Yale School of Public Health, New Haven, CT 06510, USA; 32H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA; 33Department of Preventive Medicine, College of Medicine, University of Hong Kong, Hong Kong SAR, People’s Republic of China; 34Department of Pathology, Li Ka Shing Faculty of Medicine, University of Hong Kong, Hong Kong SAR, People’s Republic of China; 35Department of Occupational Medicine and Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Huazhong University of Science and Technology, Wuhan 430073, People’s Republic of China; 36Department of Preventive Medicine, University of Hawaii, Honolulu, HI 96813, USA; 37Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; 38Department of Etiology & Cancerogenesis, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, People’s Republic of China; 39State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, People’s Republic of China; 40Department of Preventive Medicine, College of Medicine, National Taiwan University, Taipei 10617, Taiwan, ROC; 41Department of Pathology, University of Hong Kong, Hong Kong SAR, People’s Republic of China; 42Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; 43Department of Preventive Medicine, College of Medicine, National Taiwan University, Taipei 10617, Taiwan, ROC; 44Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; 45Department of Preventive Medicine, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; 46Department of Preventive Medicine, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; 47Department of Pathology, University of Hong Kong, Hong Kong SAR, People’s Republic of China; 48Department of Thoracic Surgery, School of Medicine, Vanderbilt University, Nashville, TN 37232, USA; 49Department of Preventive Medicine, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; 50Department of Preventive Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; 51International Epidemiology Institute, Rockville, MD 20850, USA.

The American Journal of Human Genetics 96, 487–497, March 5, 2015 487
Detectable mosaicism is the presence of two or more genetically distinct populations of cells in an individual who has developed from a single zygote. The clonal expansion of acquired post-zygotic mutations, such as large-scale gains, losses, and copy-neutral uniparental disomy, can result in the co-existence of aberrant cellular populations with normal germline DNA. Clonal mosaicism can also contribute to diverse phenotypes depending on developmental timing, the tissue involved, the genomic location of the mutation, and the percentage of cellular populations affected. Compared to constitutional defects in the same regions, mosaic abnormalities can result in milder phenotypes, as observed for neurofibromatosis type 1 (MIM 162200) and trisomy 21 (MIM 190685); interestingly, these same mutations have been observed in apparently healthy individuals. A spectrum of clinical phenotypes, including Maffucci syndrome (MIM 614569), McCune-Albright syndrome (MIM 174800), nevus sebaceous (MIM 162900), Ollier disease (MIM 166000), Proteus syndrome (MIM 176920), and mosaic RASopathies, have been associated with mosaicism.

Until recently, estimates of the rates of human mosaicism involving large structural events were unavailable. Early evidence demonstrated somatic mosaicism in monozygotic twins and differentiated human tissues but provided no estimates of rates in human populations. The combination of large datasets and improved methodology for analysis of genome-wide SNP microarray data has enabled genome-wide surveys of large structural mosaic events in blood and buccal DNA. An initial population-based case-control genome-wide association study (GWAS) of 1,991 individuals with bladder cancer reported autosomal mosaic abnormalities (e.g., structural events >2 Mb) in blood or buccal DNA from 1.7% of the overall study sample. Subsequent analyses of a larger set of GWASs involving 57,853 individuals, described as the total GWAS set I (TGSI), presented evidence that clonal mosaicism was strongly associated with greater age and weakly associated with male gender and overall solid-tumor risk, in particular lung and kidney cancer. A concurrently published study, involving 50,222 individuals, from the Gene-Environment Association Studies (GENEVA) Consortium observed an association between mosaicism and age, although no significant associations were observed with gender or solid tumors. An additional study by Forsberg and colleagues detected age-related structural changes in leukocyte DNA from paired monozygotic twins and single-born subjects in 3.4% of individuals aged 60 years or older but failed to detect mosaic events in individuals aged 55 years or younger. Other studies have subsequently confirmed the presence of detectable autosomal mosaicism in older populations, as well as demonstrated an age-specific relationship with mosaicism on the Y chromosome.

Furthermore, recent evidence indicates that somatic mosaicism might be an important contributor to unexpected familial recurrences of genomic disorders.

We confirmed the presence of clonal mosaic events greater than 2 Mb in an independent sample set of cancer-affected individuals and control individuals and conducted a meta-analysis of the events from our sample set and the two prior investigations to refine our understanding of the landscape of events. Study subjects from our new sample set, hereafter referred to as the total GWAS set II (TGSI), were drawn from published GWASs investigating cancer-susceptibility risk and were analyzed in the Cancer Genomics Research Laboratory of the National Cancer Genomics Research Laboratory of the National Cancer
mosaic events, we first combined TGSI 20 and TGSII, both TGSI and TGSII. Common modifications included adjust-
netic identity was consistent across duplicate samples. (LRR) and B allele frequency (BAF), respectively, and (4) ge-
clonal mosaic regions by identifying breakpoints in B dev-
detection analysis) was used to detect clonal mosaic events
software package R-GADA (based on genome alteration
algorithms incorporated into the
omeStudio software suffer biases from assay chemistry
plex ligation-dependent probe amplification, and fluores-
der techniques (e.g., single tandem repeat, multi-
mosaic copy gain, whereas values less than 0 indicate
of the ratio of observed total signal intensities to expected
signal intensities for a SNP. LRR values greater than 0 indi-
cate mosaic copy gain, whereas values less than 0 indicate
loss. The BAF, a measure of allelic imbalance, is calculated
as the ratio of signal intensity between two alleles at each
SNP in relation to estimated genotype clusters; it is thus
a calculation of the frequency of the B allele for a biallelic
SNP with alleles A and B. BAF values for heterozygous SNPs
that deviate from 0.5 are indicative of mosaic copy-num-
ber changes or copy-neutral changes associated with ac-
quired uniparental disomy.

To assess copy-number changes and allelic imbalances, we estimated LRR and BAF. LRR provides a metric for as-
sessing copy-number change via the calculation of log_2 of the ratio of observed total signal intensities to expected
signal intensities. A value greater than 0 indicates a mosaic copy number change, while values less than 0 indicate
loss. The BAF, a measure of allelic imbalance, is calculated as the ratio of signal intensity between two alleles at each
SNP in relation to estimated genotype clusters; it is thus
a calculation of the frequency of the B allele for a biallelic
SNP with alleles A and B. BAF values for heterozygous SNPs
that deviate from 0.5 are indicative of mosaic copy-number changes or copy-neutral changes associated with ac-
quired uniparental disomy.

To improve accuracy, we used a quantile normalization
approach similar to that used by Staaf et al. and Diskin
et al. because LRR and BAF estimates from Illumina Ge-
nomeStudio software suffer biases from assay chemistry
and DNA concentration. For TGSI and TGSII, a mosaic-
tation-alteration-detection algorithm incorporated into the
software package R-GADA (based on genome alteration
detection analysis) was used to detect clonal mosaic events
greater than 2 Mb in size from corrected LRR and BAF
values. The GADA segmentation algorithm detected clonal mosaic regions by identifying breakpoints in B_ser-
values with the use of sparse Bayesian learning and back-
ward elimination. The assigned event type was based on the mean LRR value, and the mosaic proportion of
abnormal cells was estimated from the BAF values. The
detection algorithm has been previously validated with
laboratory techniques (e.g., single tandem repeat, multi-
plex ligation-dependent probe amplification, and fluores-
cent in situ hybridization) and is described in greater
detail in the methods and supplementary material of the
original Jacobs et al. analysis. Investigators blind to study
outcomes conducted a manual review to confirm events in
TGSI and TGSII. Common modifications included adjusting
boundaries of mosaic events or merging or separating
adjacent events.

To improve our understanding of the landscape of clonal
mosaic events, we first combined TGSI and TGSII, both
of which were analyzed with the identical pipeline
described above. Subsequently, we extracted events >2 Mb from the Laurie et al. study (GENEVA). GENEVA used
a circular-binary-segmentation algorithm with the
Bioconductor packages DNACopy and GWAStats. The
overall reproducibility between the GENEVA algorithm
and TGSI algorithm in detecting events >2 Mb was as-
assessed in 5,510 lung cancer samples included in both
GENEVA and TGSI. The comparability in called events
>2 Mb was 75%; 83 of the total 111 events were detected
by both algorithms, 20 were detected by TGSI only, and
8 were detected by GENEVA only. The 5,510 replicated
lung cancer samples and 235 individuals with inadequate
consent were removed from the GENEVA sample in the an-
alyzes described herein. For more details on the detection
methods, please see the methods and supplementary ma-
terials in the previously published TGSI and GENEVA
reports, which include laboratory confirmation of select
samples in TGSI.

In the 24,849 individuals in TGSII, our method detected
341 clonal mosaic events across 255 autosomes in 168 in-
dividuals (Table S1). We detected 69 events (20%) with
mosaic copy gain, 90 events (26%) with mosaic copy
loss, and 163 events (48%) with copy-neutral acquired uni-
parental disomy. 19 events (6%) were complex in nature
and not amenable to distinct classification.

Detected events from our TGSII were combined with
TGSI, and then a meta-analysis with GENEVA was per-
formed on a sample set of a total of 127,179 individuals,
whom 1,315 mosaic events were detected in 925 partici-
pants; the overall rate of individuals with detected mosaic
ism was 0.73% (95% confidence interval [CI] = 0.68–
0.78%). Of the 925 participants with detected events,
797 (86%) had only one event, and 128 (14%) harbored
multiple mosaic events. Compared to the Poisson expecta-
tion, which is that only seven individuals have multiple
events, a highly significant excess of individuals with mul-
tiple events was observed (p value = 6.5 × 10^{-36}).
Although complex rearrangements affecting multiple re-
regions of the genome could partially account for the excess,
differences in mosaic proportion of events in some individ-
uals suggest more complex mechanisms over long periods
of time as well.

Approximately half of the detected events were mosaic
copy-neutral uniparental disomy (48%), followed by
mosaic losses (34%) and mosaic gains (17%). To charac-
terize genomic location and potential recurring events,
we generated Circos plots. The majority of mosaic gains
were observed on chromosomes 8, 12, and 15, the majority
of mosaic losses were observed on chromosomes 13 and
20, and mosaic copy-neutral events were primarily
observed on chromosomes 9 and 14 (Figure 1). Detected
mosaic events clustered regionally on chromosomal arms
on the basis of their copy-number state (Table 1). Mosaic
copy-neutral events occurred primarily on the telomeric
ends of chromosomes; 33% included the p telomere, and
54% included the q telomere. Mosaic losses were
commonly observed in interstitial chromosomal regions that did not involve telomeres or centromeres. Of all of the copy-number states, mosaic gains most commonly involved whole chromosomes. Only 1.7% of events were interstitial and spanned the centromere, suggesting that for interstitial mosaic events, involvement of a centromere could be uncommon.

It is notable that losses at certain regions, such as 20q (chr20: 40,425,000–42,155,000; UCSC Human Genome Browser hg18) and 13q14 (chr13: 49,590,000–49,983,100; UCSC hg18), are observed in leukemias (e.g., myelogenous and lymphoblastic leukemia, respectively). Clustering of losses in these regions suggests that these events are non-random. Moreover, aggregation of mosaic-event locations on chromosomal arms by copy-number state suggests common mechanisms. For example, copy-neutral telomeric events could be due to mitotic recombination followed by clonal expansion. Breakpoint analyses of regions surrounding mosaic events might aid in understanding mechanisms responsible for event initiation, but the current resolution of event boundaries in SNP microarrays is limited as a result of insufficient probe density. Further work is required to investigate the different types of events that could lead to large structural mosaicism.

Table 1. Distribution of Mosaic Copy-Number State by Chromosomal-Arm Location

<table>
<thead>
<tr>
<th></th>
<th>Gain</th>
<th>Neutral</th>
<th>Loss</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole chromosome</td>
<td>60</td>
<td>(65%, 27%)</td>
<td>25</td>
<td>(27%, 4%)</td>
</tr>
<tr>
<td>Telomeric p</td>
<td>18</td>
<td>(7%, 8%)</td>
<td>205</td>
<td>(78%, 33%)</td>
</tr>
<tr>
<td>Telomeric q</td>
<td>67</td>
<td>(15%, 30%)</td>
<td>338</td>
<td>(75%, 54%)</td>
</tr>
<tr>
<td>Interstitial</td>
<td>73</td>
<td>(16%, 33%)</td>
<td>51</td>
<td>(11%, 8%)</td>
</tr>
<tr>
<td>Spans centromere</td>
<td>5</td>
<td>(22%, 2%)</td>
<td>10</td>
<td>(43%, 2%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>223</td>
<td>(17%, 100%)</td>
<td>629</td>
<td>(49%, 100%)</td>
</tr>
</tbody>
</table>

Events classified as “whole chromosome” from the combined dataset span an entire chromosome, “telomeric p” and “telomeric q” are events that include the telomere on the p or q arm, respectively, “interstitial” events do not include a telomere or centromere, and “spans centromere” indicates interstitial events that overlap a chromosome’s centromere. Event counts are indicated in parentheses (row percent, column percent).

To identify characteristics associated with increased risk of large clonal mosaic events, we evaluated age, gender, ancestry, and cancer status. Age at time of DNA collection was available for all GENEVA and TGSI participants, but for TGSI participants, the date of diagnosis (for cancer subjects) or the age at the time of participation (for control subjects) was substituted when age at time of DNA collection was missing. Categorical variables were constructed for the 5-year age groups of 50–54, 55–59, 60–64, 65–69, and 70–74 and for 75 years or older, whereas individuals under 50 years of age were considered the reference group. We used reference populations from the HapMap project and the GLU (Genotype Library and Utilities) software package to estimate continental-ancestry proportions for each individual. Terms were fit for percentage of African and Asian ancestry, whereas European ancestry served as the referent. Indicator variables were used for adjusting analyses for effects related to the individual contributing studies. Sensitivity analyses using mixed models, case-control matching, and pooled analyses were also used to investigate the robustness of statistically significant findings. All statistical analyses were performed in R version 3.0.1 on a 64-bit Unix platform.

Increasing age is the variable most strongly associated with clonal mosaicism (Figure 2). In logistic regression
models adjusted for gender, ancestry, cancer subtype, and contributing study, highly significant age effects were observed in both a meta-analysis and a pooled analysis of GENEVA, TGSI, and TGSII (Figure 3A). The effect of the 5-year age groups was significantly associated with detectable clonal mosaicism overall (p value = 5.5 × 10^{-31}), showed no evidence of heterogeneity across the study (p value = 0.71), and remained significant when the analysis was restricted to a subset of cancer-free control individuals (p value = 8.92 × 10^{-12}). Compared to individuals under 50 years old, individuals aged 75 years or older had an approximate 6-fold increase in detection of large-scale mosaic events (p value = 2.22 × 10^{-16}, 95% CI = 4.16–10.09).

Our meta-analysis strengthens the robust age association previously observed in GENEVA and TGSI. Although the inclusion of age at diagnosis as a substitute for age at the time of DNA collection could have introduced measurement error in individual TGSII participants, the TGSII and overall pooled age association agree with the original age estimates from the GENEVA and TGSI studies. Although an association with age was observed, it is important to note that our analysis does not provide insight into whether the events were generated early in life and later positively selected by rapid expansion of a second clonal population or generated later in life as a result of decreased cellular diversity and senescence.

The effect of gender was also significant. Mosaic events were more frequently observed in males than in females (0.98% versus 0.56%, respectively; Figure 3B). Removing sex-specific cancers (e.g., endometrial and prostate cancers) and adjusting for ancestry, 5-year age group, cancer subtype, and contributing study, we found a significant association with male gender (odds ratio [OR] = 1.44, 95% CI = 1.15–1.81, p value = 0.002) and no evidence of heterogeneity of effect across the study (p value = 0.88). In the 37,942 cancer-free control individuals, the

<table>
<thead>
<tr>
<th>Age Effect</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GENEVA</td>
<td>7.59</td>
<td>[4.00, 14.40]</td>
</tr>
<tr>
<td>TGSI</td>
<td>5.79</td>
<td>[3.37, 9.93]</td>
</tr>
<tr>
<td>TGSII</td>
<td>5.00</td>
<td>[2.11, 11.86]</td>
</tr>
<tr>
<td>Meta</td>
<td>6.17</td>
<td>[4.25, 8.96]</td>
</tr>
<tr>
<td>Pooled</td>
<td>6.48</td>
<td>[4.16, 10.09]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male Effect</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GENEVA</td>
<td>1.35</td>
<td>[0.56, 3.24]</td>
</tr>
<tr>
<td>TGSI</td>
<td>1.42</td>
<td>[1.11, 1.81]</td>
</tr>
<tr>
<td>TGSII</td>
<td>2.41</td>
<td>[0.30, 19.45]</td>
</tr>
<tr>
<td>Meta</td>
<td>1.42</td>
<td>[1.13, 1.79]</td>
</tr>
<tr>
<td>Pooled</td>
<td>1.44</td>
<td>[1.15, 1.81]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>African Ancestry Effect</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GENEVA</td>
<td>0.71</td>
<td>[0.21, 2.43]</td>
</tr>
<tr>
<td>TGSI</td>
<td>0.39</td>
<td>[0.19, 0.77]</td>
</tr>
<tr>
<td>TGSII</td>
<td>0.96</td>
<td>[0.02, 55.25]</td>
</tr>
<tr>
<td>Meta</td>
<td>0.45</td>
<td>[0.25, 0.83]</td>
</tr>
<tr>
<td>Pooled</td>
<td>0.38</td>
<td>[0.21, 0.71]</td>
</tr>
</tbody>
</table>

Figure 3. Forest Plots of Associations with Clonal Mosaicism
Associations between clonal mosaicism and (A) age-group, (B) gender, and (C) ancestry.
gender association was of similar magnitude but marginally not significant (OR = 1.39, p value = 0.06), most likely as a result of the reduced number of individuals with events (n = 303). The elevated rates of mosaic events observed in males could be partially attributable to higher male-specific rates of hematologic malignancies, a set of malignancies previously found to be associated with clonal mosaicism. Further work will be required for understanding the scope and implications of this observation.

An association between genotype-inferred ancestry (defined as the percentage of continental origin from Africa, Asia, or Europe) and clonal mosaic events was also evident (Figure 3C). Logistic regression analyses adjusted for gender, 5-year age group, cancer subtype, and study indicated that relative to individuals of European ancestry, individuals of African ancestry were at a reduced risk (OR = 0.38, 95% CI = 0.21–0.71, p value = 0.003). Heterogeneity testing detected no heterogeneity (p value = 0.66), and the effect retained significance when the analysis was restricted to cancer-free control individuals (OR = 0.23, 95% CI = 0.09–0.60, p value = 0.003). No significant difference was observed between Asian and European ancestry. Mechanisms relating to ancestry-specific differences in rates of clonal mosaicism are poorly understood, and further work is needed for better understanding this relationship.

We further investigated the associations between clonal mosaicism and risk of solid (non-hematological) tumors overall and of tumor subtypes as per the previous TGSI finding. Analyses adjusted for gender, ancestry, 5-year age group, and contributing study indicated that solid tumors were associated with clonal mosaicism in blood or buccal tissue (OR = 1.29, 95% CI = 1.11–1.50, p value = 8.1 × 10⁻⁴). Additional cancer-specific analyses were performed for all solid-tumor subtypes present in our study (Table 2). Endometrial, kidney, liver, lung, and prostate cancers suggested preliminary evidence of significant associations with clonal mosaicism; however, only lung cancer maintained statistical significance after correction for multiple testing. Interestingly, all solid-tumor subtypes showing preliminary evidence of an association with mosaicism had ORs greater than 1, but the sample sizes per tumor were small, and the estimates were unstable for the risk effects measured. Circos plots showing mosaic-event location for cancer-free control individuals and for individuals with solid tumors are displayed in Figures 1B and 1C. No differences in event clustering or copy-number state were observed among the individuals with solid tumors. Lung cancer was the primary contributor to the overall cancer association, and it is notable that lung cancer is a smoking-related cancer; however, previous data suggest that smoking is not significantly associated with autosomal mosaicism in blood tissue. Potential mechanisms linking clonal mosaicism in blood to solid-tumor risk might include poor overall genome maintenance unable to repair genomic alterations or immunologic dysfunction in mosaic immune cells and the subsequent poor clearance of pre-cancerous cells from solid tissues. Further studies are required to determine whether detectable clonal mosaicism could be a useful biomarker in screening individuals for increased risk of developing solid tumors. This is in

<p>| Table 2. Solid-Tumor Associations with Clonal Mosaicism |</p>
<table>
<thead>
<tr>
<th>n</th>
<th>OR</th>
<th>95% CI</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>46,831</td>
<td>1.29</td>
<td>1.11–1.50</td>
</tr>
<tr>
<td>Bladder</td>
<td>4,995</td>
<td>1.29</td>
<td>0.90–1.85</td>
</tr>
<tr>
<td>Breast</td>
<td>2,814</td>
<td>1.06</td>
<td>0.53–2.10</td>
</tr>
<tr>
<td>Endometrium</td>
<td>872</td>
<td>2.66</td>
<td>1.16–6.12</td>
</tr>
<tr>
<td>Esophagus</td>
<td>1,910</td>
<td>0.89</td>
<td>0.34–2.34</td>
</tr>
<tr>
<td>Glioma</td>
<td>1,729</td>
<td>0.78</td>
<td>0.30–2.06</td>
</tr>
<tr>
<td>Kidney</td>
<td>1,565</td>
<td>1.81</td>
<td>1.06–3.11</td>
</tr>
<tr>
<td>Liver</td>
<td>13</td>
<td>10.40</td>
<td>1.32–81.9</td>
</tr>
<tr>
<td>Lung</td>
<td>13,015</td>
<td>1.54</td>
<td>1.21–1.97</td>
</tr>
<tr>
<td>Ovary</td>
<td>543</td>
<td>2.60</td>
<td>0.93–7.28</td>
</tr>
<tr>
<td>Pancreas</td>
<td>3,923</td>
<td>0.83</td>
<td>0.53–1.29</td>
</tr>
<tr>
<td>Prostate</td>
<td>10,456</td>
<td>1.27</td>
<td>1.00–1.60</td>
</tr>
<tr>
<td>Skin</td>
<td>1,949</td>
<td>0.96</td>
<td>0.43–2.13</td>
</tr>
<tr>
<td>Stomach</td>
<td>2,278</td>
<td>1.52</td>
<td>0.70–3.28</td>
</tr>
<tr>
<td>Testis</td>
<td>649</td>
<td>1.77</td>
<td>0.24–13.0</td>
</tr>
</tbody>
</table>

Combined analysis of solid-tumor associations adjusted for gender, ancestry, 5-year age group, and contributing study. “n” denotes total sample size. ORs and 95% CIs are reported for solid tumors overall and by cancer subtype.
contrast to hematological malignancies, in which detectable mosaicism in blood could be an early indicator of leukemic and pre-leukemic clones.

An association between rate and event size was noted: smaller events were observed to have higher rates than larger events. An inverse exponential relationship fit the trend well ($R^2 = 0.81$). This association was also present in the substrata of mosaic gains, copy-neutral events, and losses. Event size was also investigated across copy-number states. With a median event size of 60.8 Mb, mosaic copy gains, on average, were largest. Mosaic copy-neutral events and mosaic losses had median sizes of 39.8 and 17.0 Mb, respectively. The larger average size of mosaic gains and copy-neutral events might highlight the detrimental nature of induced monosomy and reduced copy number of mosaic losses and provide insight into understanding which mutational events undergo clonal selection. By restricting our analysis to events larger than 2 Mb in size to reduce the false-positive rate of our detection algorithm, we most likely missed many smaller mosaic events, from mosaic point mutations to events several kilobases in size. Extrapolating our observations to smaller events suggests that the copy-number distribution of these events is most likely skewed toward mosaic losses. This could prove to have great importance in both disease risk and heterogeneity of disease phenotypes. Although the association between event rate and size was present in every strata of copy-number state, the overall association might be skewed by the abundance of small copy losses in likely driver regions of hematologic cancers, such as losses at 13q14 and 20q. Further refinements in genotyping technologies and detection algorithms targeted at detecting smaller mosaic events in next-generation sequencing should refine our understanding of the landscape of detectable clonal mosaicism.

Mosaic proportion, namely, the percentage of cells with large structural events that differ from germline DNA, was distinct across event copy-number state. Mosaic copy gains and copy losses were observed, on average, to have higher mosaic proportions (medians of 0.35 and 0.33, respectively), whereas mosaic copy-neutral events had a lower median proportion of 0.19. The range of the mosaic proportions provides insight into the detectable range of our methodology. Copy-neutral mosaic events can be detected in the range 0.06–0.95, whereas mosaic copy gains can be detected in the range 0.10–0.90.

The relationship between age and mosaic proportion was also investigated. A significant positive association between age and mosaic proportion was observed for copy-neutral events (p value $= 1.2 \times 10^{-4}$), suggesting that on a population level, mosaic proportion...
might gradually increase with age. To further assess the evolution of mosaic events over time, we analyzed serial samples from the Prostate, Lung, Colon, and Ovary Prevention Trial (PLCO). Two to four DNA samples collected at least 1 year apart were analyzed for PLCO individuals with detectable clonal mosaicism in TGSI and TGSII. In total, we tracked large structural mosaic events for 58 autosomes (Figure 7) detected in 47 individuals. The set included 24 events with mosaic loss, 27 with copy-neutral loss of heterozygosity, and 7 with gain. Although examples existed where events had stable or decreasing mosaic proportion over time, most events were observed to increase in mosaic proportion. Fitting a linear mixed model with a zero intercept and a random effect for each event, our analysis suggests that with each year increase in age, the overall fraction of mosaic proportion increases on average by approximately 1.44% (p value \(3.3 \times 10^{-7}\)). Significant increases in mean mosaic proportion were seen over time in the strata of mosaic losses and mosaic copy-neutral events, but because of limited sample size (n = 7), the increase was not significant for mosaic gains. Together, these observations suggest that most detectable mosaic events confer some form of selective advantage that enables cellular clones to increase in frequency over time in relation to cells with normal karyotypes.

A few limitations relating to the available data are worth considering. The studies used for analysis were primarily designed as GWASs of cancer (GENEVA, TGSI, and TGSII) and other traits (GENEVA), and study participants were drawn from cohort and case-control studies. Rate estimates of mosaicism, as provided by the combination of individuals included in these studies, could imperfectly represent underlying population prevalence. However, the consistency of effect estimates across GENEVA, TGSI, and TGSII suggests that our findings are robust. Additionally, although incomplete, adjustment for factors such as age group, gender, ancestry, cancer subtype, and contributing study minimized confounding effects. Because additional bioinformatics methods are needed to detect mosaicism on the sex chromosomes and because there is poor commercial array coverage of the Y chromosome, this report focused on detectable autosomal mosaicism. Another group has reported a similar association between Y mosaicism and increasing age.25
Our analysis is distinctive in that it involves a meta-analysis of 127,179 individuals and thus provides a more comprehensive portrait of the landscape of large structural mosaic genetic alterations. Our meta-analysis allowed for a more precise age effect, strengthened prior evidence of a gender effect, and found evidence of an ancestry effect. Additionally, we found evidence indicating that the number of mosaic events increases as event size decreases, suggesting that our detected events might represent the tip of the iceberg in relation to many smaller mosaic events that most likely exist and are currently undetected. Furthermore, we were able to analyze serial samples over multiple collection time points to show an overall increase in mosaic percentage as age increases.

The investigation of human clonal mosaicism can provide new insights into aging, as well as shed light on possible precursors of disease. Our results suggest that genome maintenance, particularly in relation to aging, could have pleiotropic consequences, although it is not clear whether all mosaic events are necessarily deleterious. The long-held assumption that germline DNA remains static during the course of life is under reconsideration, which is particularly important for the comparative study of cancer genomes. As the detection of smaller mosaic events across tissue types improves, we can refine our understanding of the landscape of mosaic events across the spectrum of genetic events. In turn, this could represent an important step forward in the investigation of how our once “stable” germline DNA might slowly erode into a complex mosaic over time and contribute to disease heterogeneity.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments and one table and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.01.011.

Acknowledgments

Acknowledgment of funding sources is available in the Supplemental Data. B.R.-S. and L.A.P.-J. are currently employee and scientific advisor, respectively, of qGenomics. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the NIH.

Received: October 7, 2014
Accepted: January 12, 2015
Published: March 5, 2015

Web Resources

The URLs for data provided herein are as follows:

Genotype Library and Utilities (GLU), https://code.google.com/p/glu-genetics/
UCSC Genome Browser, http://genome.ucsc.edu

References


